Computational Social Science Project Report

Student name: Nguyen Xuan Binh
Student ID: 887799

Table of contents

Task A: Review a related paper (100 points)

Review the paper – Women through the glass ceiling: gender asymmetries in Wikipedia
https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-016-0066-4

Please read the paper carefully and write 1-2 paragraphs each to summarize the following in your own words:

- **Primary goal of the paper (30 points)**
- **Main methods (20 points)**
- **Key findings (30 points)**
- **Brief description of one follow-up analysis you could have done based on the paper if you had unlimited access to data and compute resources (20 points)**

1. Primary goal of the paper (30 points)

The first primary goal of this paper is to investigate the possible underlying gender and linguistic bias against women on Wikipedia pages, a global encyclopedia on various topics. Particularly, this paper's research is limited to personal biographies of men and women in the English language editions because this language has the most pages, and other languages have nearly 100% overlapping pages with English. The paper claims that the editor community has a narrow diversity, with a majority of white male editors, so there should exist a potential bias in the content of the biographies. As a result, the paper tries to investigate the properties of Wikipedia articles: notability, topical focus, linguistic bias, structural

properties, and meta-data presentation. Based on these aspects, they can compare if articles about men and women have nearly similar structures or whether there is an inherent difference based on gender only.

The paper proposes a concept of the glass-ceiling effect, which refers to the situation in which women cannot reach higher positions due to an invisible barrier of gender bias. It hypothesizes that if the entry point of Wikipedia functions as a glass ceiling, fewer women will be included and those included will be more notable than their male counterparts on average. Based on its findings about the glass ceiling effect and biased linguistic uses against women, this paper concludes that there is indeed a gender inequality in Wikipedia articles. As such, the second primary goal of the paper is to promote Wikipedia's internal policies for better gender equality representation. The paper expects its work can contribute to increased awareness about gender biases online. It also suggests that Wikipedia editors consider revising its guidelines to account for women's low visibility and encourage a less biased use of language.

2. Main methods (20 points)

To study gender bias in Wikipedia, the authors use The DBpedia dataset and split it into pre-1900 and post-1900. The pre-1900 sample contains all people born before 1900, while the post-1900 sample consists of people born in or after 1900. The main methods this paper uses to assess the extent to which gender bias manifests in Wikipedia are the comparisons between articles about men and women along the following dimensions: global notability according to external and internal proxy measures, topical focus, linguistic bias, structural properties like meta-data and hyperlink networks. Furthermore, these comparisons are applied for both pre and post-1900 datasets since the authors believe the advent of the Internet significantly affects the varying level of bias in Wikipedia articles.

In detail, this paper uses the number of language editions as an internal measure (Wikipedian editors) and Google search volume as an external measure to quantify public interest. For topical bias, the paper analyzes three topics that could be over-represented in articles about women: gender, relationship, and family. For linguistic bias, the paper uses a lexicon-based approach and syntactic annotations to detect abstract and subjective language. For structural properties, this paper relies on the search result rankings, which are often informed by centrality measures such as PageRank. For meta-data, it compares the relative proportions of attribute presence between genders using chi-square tests. Finally, the paper computes the PageRank of articles about people to estimate the hyperlink networks.

3. Key findings (30 points)

The findings this paper claims are based on the proposed speculations in its methodologies. Regarding inequalities in global notability thresholds, this paper found that the gap between men and women is larger for people with low or medium levels of global notability than for global stars. This paper explains the high men-to-women ratio for local heroes: men love to create articles about themselves because they are more self-absorbed than women, or more information may be available online about less notable men than about them less notable women. However, this paper is hesitant to claim the glass ceiling effect for inclusion in Wikipedia of women born before 1900 because the more historic a person is, the more notable they are globally. Conversely, young people born recently may benefit from the availability of digital information, making them more likely to be recognized by editors. Another critical finding about Google search trends is that women in Wikipedia are slightly more of interest to the world.

Regarding topical and linguistic asymmetries, this paper finds that besides professional and topical areas, words in the gender, relationship, and family categories are more dominant in articles about women born before 1900 but less pronounced in articles about people born since 1900. Overall, women tend to have more words related to family, gender, and relationships than men. Additionally, adjectives are much more likely to describe the positive aspects of men's biographies and the negative aspects of women's biographies. Regarding structural inequalities, many attributes such as activeYearsEndDate, careerStation, and position are more frequently used to describe men. All of these attributes are related to sports due to the prominence of men in sports-related. Attributes deathDate, and deathYear are more frequently used for men born before 1900 because women's life was less well documented back then. In contrast, attribute birthName and spouse is more frequently used for women in recent times because married women change their surnames to those of their husbands in some cultures. Furthermore, women's biographies have more links to other women's articles, which stems from the reported interests of female editors in women's biographies in Wikipedia. However, a final key finding of this paper claims that the empirically observed structure of the hyperlink network puts women disadvantage and that there exists a bias in the generation of links by Wikipedia editors, favoring articles about men

4. Brief description of one follow-up analysis you could have done based on the paper if you had unlimited access to data and computing resources (20 points)

The paper mentions that their empirical results are limited to the English Wikipedia, which is biased towards western cultures. They wish to leave a more detailed exploration of gender bias across all language editions for future work. Therefore, one possible follow-up analysis that could be done based on the paper if unlimited access to data and computing resources were available is to expand the study to multiple language editions of Wikipedia or even other platforms to examine if the observed gender inequalities are consistent across different cultures and languages. This may involve applying the same methodologies in this paper to other language editions and comparing the results to those obtained in the English edition on Wikipedia. Because English puts more emphasis on the Western cultural sphere, this research results may not be true for the countries in the Eastern sphere, such as China, India, or Japan, where gender bias is harder to be quantified. Such analysis could provide insights into the extent to which gender disparities in biographies are a global phenomenon or only specific to the Western world.

For example, China is a highly secluded country from the world's digital network. Only highly notable individuals from mainland China are mentioned on Wikipedia, leaving possibly hundred thousand notable people inside China undocumented on Wikipedia. Given unlimited access to data, we can try to access Chinese databases to see whether there exists gender bias in Chinese media (although, in reality, this could be impossible). Additionally, given unlimited computing resources, we could identify language-specific issues that affect gender inequalities in each Wikipedia language edition and inform specific strategic developments to narrow the gender imbalance gap. This could provide valuable insights into the global extent of gender disparities on Wikipedia. The restricted confidentiality of data and limited resources for conducting analysis are the main constraints on this task. However, it is valuable research given the hypothesis that we have unlimited access to data and computing resources.

Task C: Additional data analysis (200 points)

For this task, please feel free to test out your own analysis ideas or focus on things you find most interesting. You can replicate ideas from the source article or develop your own. Note that the data will contain errors and missing data, so you must deal with them in your analysis. Feel free to cut corners and mention the limitations in your report. Report negative results: if you tried something that didn't work, you can still report it.

What you should submit at the end:

- One paragraph explaining what you have done, including the problem statement and brief method description (60 points)
- One paragraph highlighting your main results (140 points)
- Any figures or tables you might need to illustrate your results
- Submit your code as a pdf and make sure that understanding the results does not require reading through your code

1. Problem statement and method descriptions (60 points)

What I have tried to do in this part is to replicate the procedural idea addressed in the paper "Women through the glass ceiling: gender asymmetries in Wikipedia". Particularly, I can state my problem as **verifying whether the observations found in that paper are true or not based on my own experiment** on the Wikipedia database of articles. The criteria that I want to check are mostly suggested by the assignment descriptions. Additionally, I also add two more tasks, amounting to 8 different tasks to check various aspects of the datasets, which are the three following ones:

- politicians.json: This is processed and provided by the code template
- finns-1900-1940.json: This is filtered from politicians.json based on nationality and birth year
- artists.json: This is my own dataset, which filters the Wikipedia database by artist occupation

Methods: I mostly use the networkx library for building graphs and I did not rely on any NLP libraries for sentiment analysis. Every statistics is conducted via simple counting, sorting and summation. These are the 8 questions that I aim to answer in my project report:

1. Test if there is homophily related to gender in the politician network compared to the baseline.
2. Replicate the analysis of comparing if females/males are more likely to be "local heroes," i.e., if it is true that females are more likely to be in a larger number of language editions than males.
3. Explore how the language used to describe men and women in their summary texts differs
4. Visualize the network structure between the politicians
5. Find the most central nodes in the network and analyze how centrality is related to gender and other attributes, such as number of language editions.
6. Compare the notability of the politicians born before and after 1900 to see whether the patterns of male/female notability are similar (my own formulated task)
7. Identify the number of unique connected components of exclusive males, exclusive females or male-female mix to infer same gender between article and editors
8. Compute statistics over the artists and Finnish politicians with the same procedures in the tasks number 2 and 3 to see if the patterns are similar across profession and nationality.

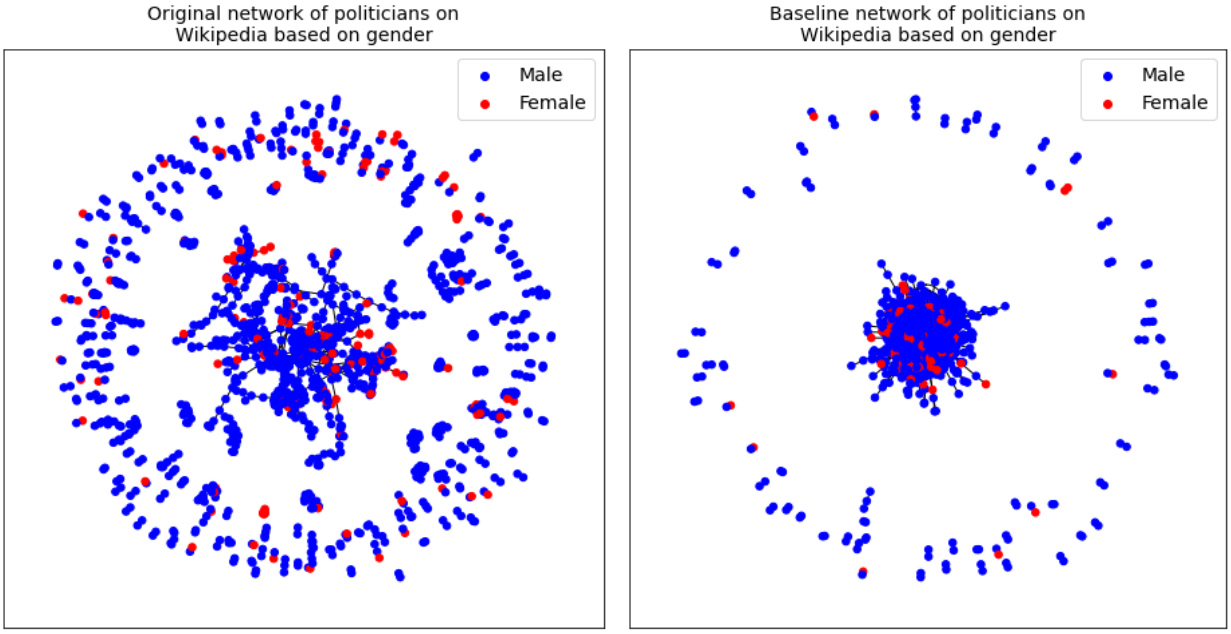
2. Highlighted results (140 points)

These are the results I have obtained from further analysis on the Wikipedia articles:

1. Homophily is the principle that contact between similar people occurs at a higher rate than among dissimilar people. It is true that the network of cross-referencing in politicians article display homophily feature, where both exclusive females and males cross-referencings are more common than the baseline network. This is my result on this task

```
Number of links between females in original network:  320
Number of links between females and males in original network:  1622
Number of links between males in original network:  3794
```

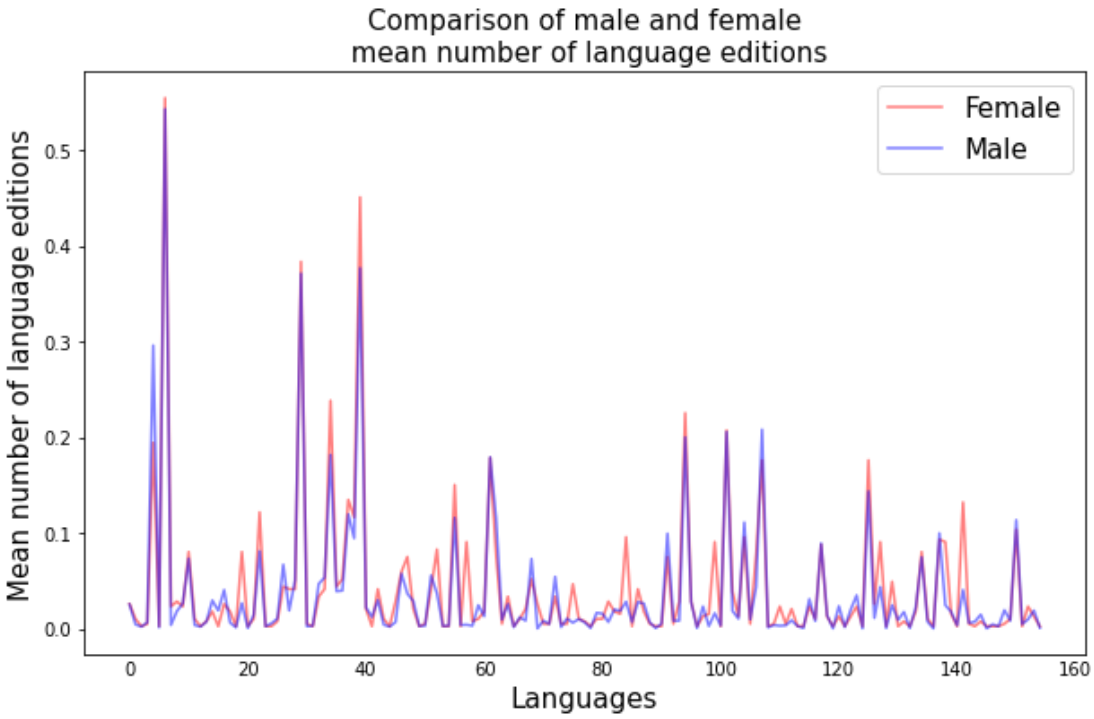
```
Number of links between females in baseline network:  255
Number of links between females and males in baseline network:  1752
Number of links between males in baseline network:  3729
```



The second graph looks like a hairball, which is relatively meaningless to interpret. However, the baseline network shows that there is a global cluster of densely connected politicians. That is, the baseline network suggests it is very likely that two people are related in some way if they are both politicians.

2. It is indeed true that females are more likely to be “local heroes” and on average more notable than their male counterparts based on the mean number of language editions and number of covering languages. This means there is an inherent glass-ceiling effect of entry point into Wikipedia for female politicians compared to males. This is my result on this task

```
Mean number of language editions for female politicians: 6.702073
Mean number of language editions for male politicians: 6.065608
Number of languages covering more female politicians than male politicians: 88
Number of languages covering more male politicians than female politicians: 67
```

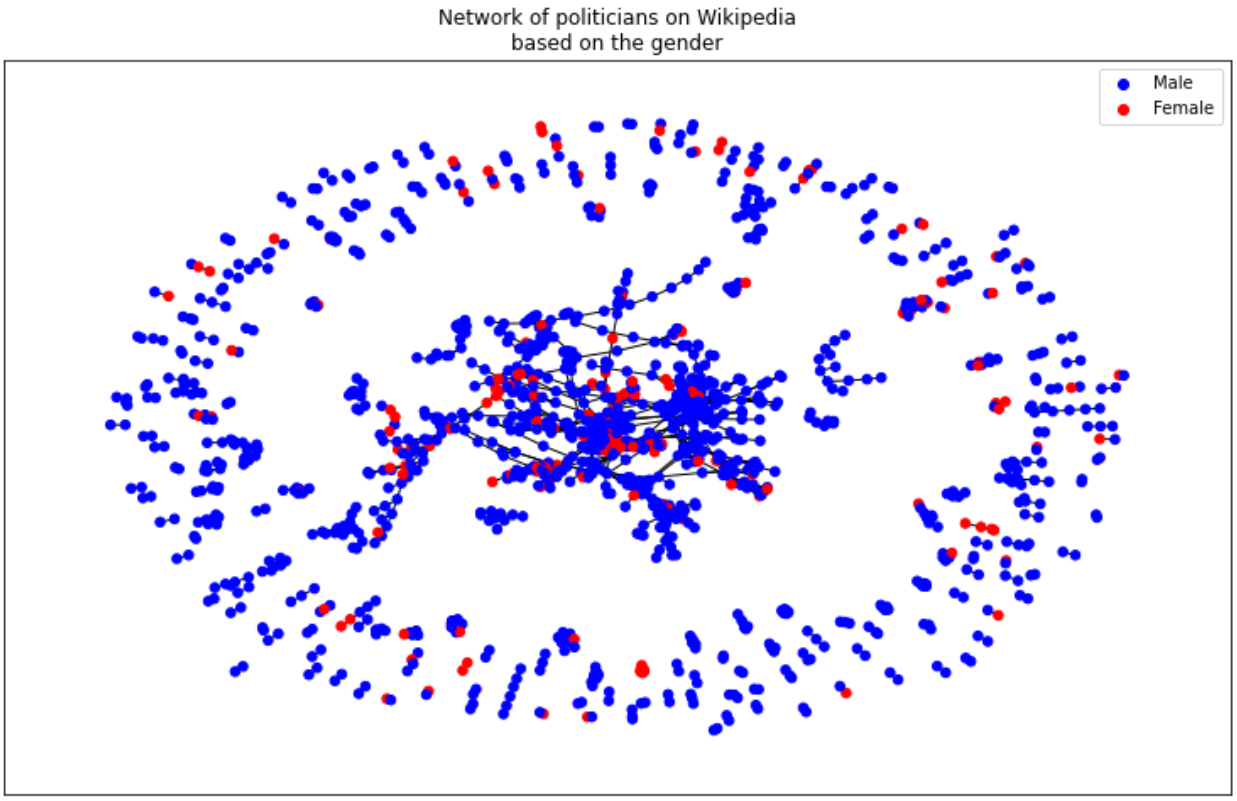


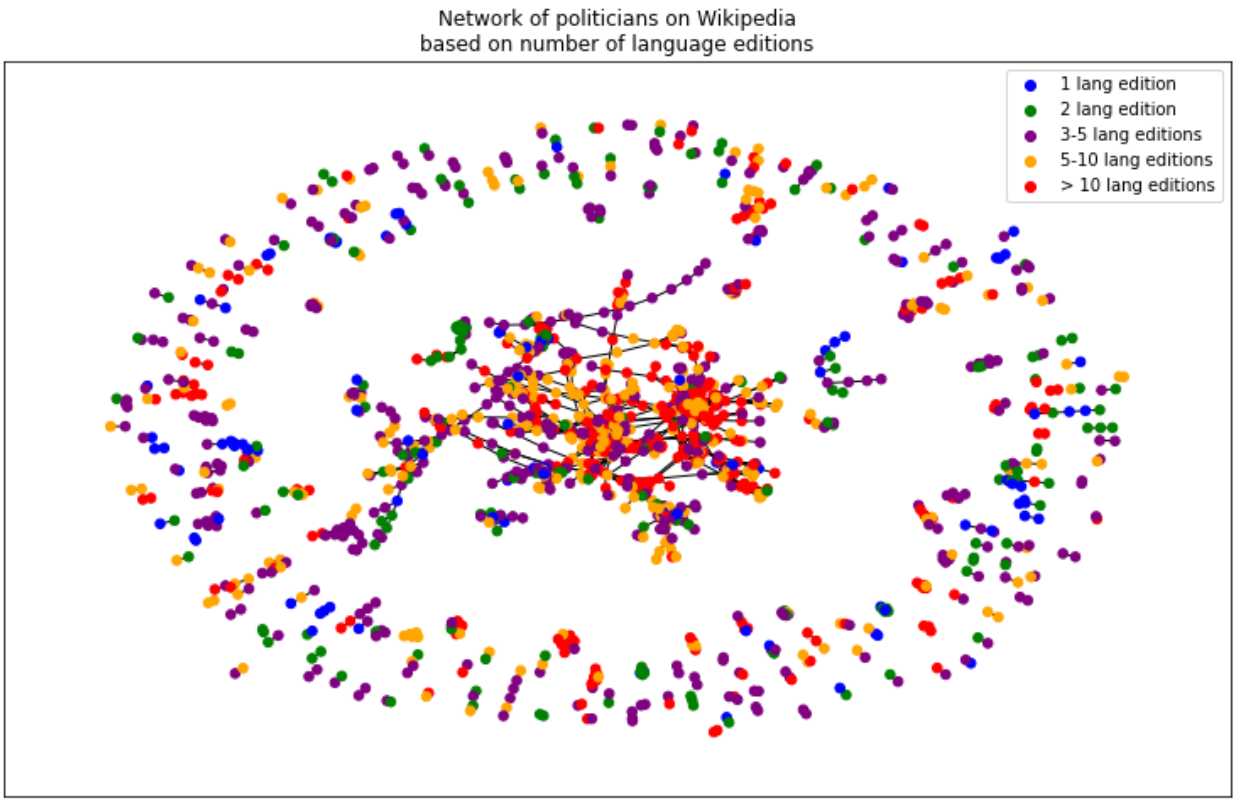
3. The rate of usage of personal life terms and professional terms appear to be equal for both males and females in my case, which contradicts the claim in the paper stating that women face biased linguistic terms related to focus on family relationships. This is my result on this task

```
Personal terms: Male: 0.7156966490299823, Female: 0.7124352331606217
Professional terms: Male: 2.4617283950617286, Female: 2.4145077720207255
```

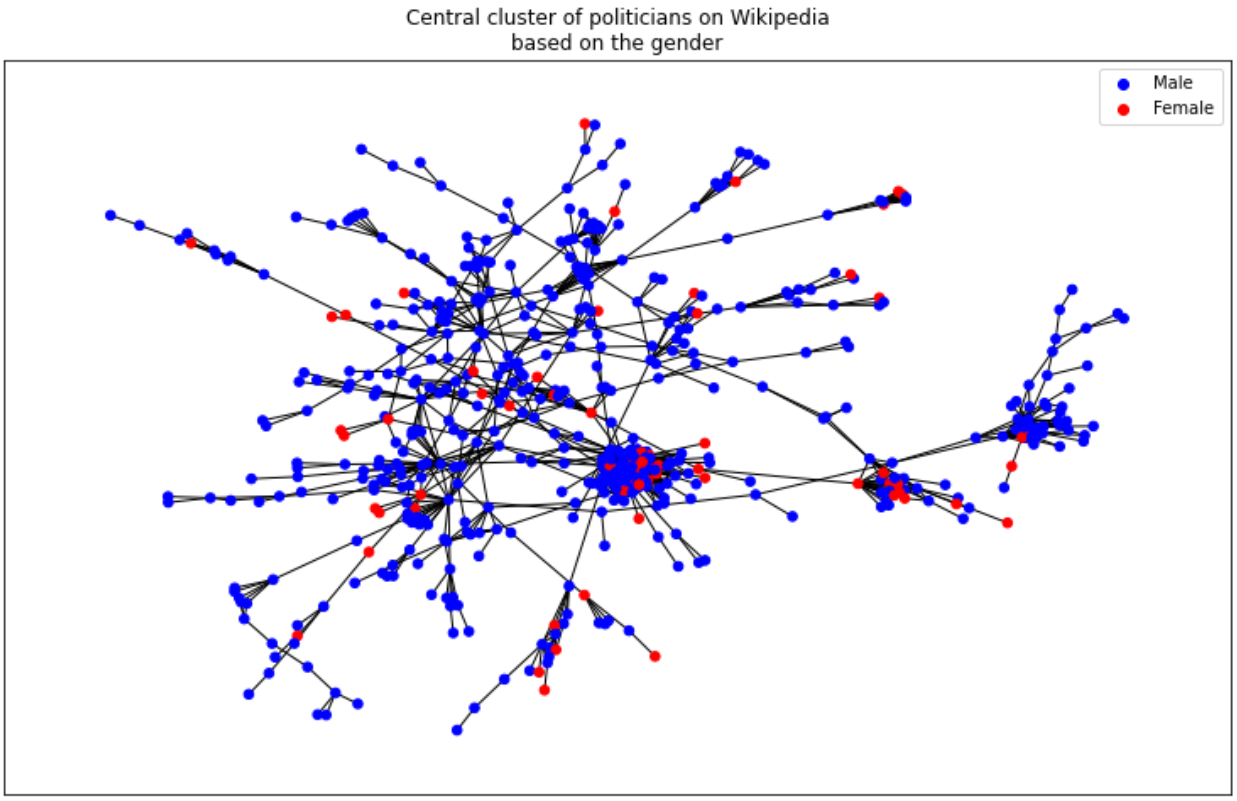
The result is unexpected because due to the vast number of politicians having no recorded genders. Had these politicians' gender been identified, the results could have been different. Another possible reason is the paper uses the whole article contents, while I only depend on the summary field of each politician. This is a possible shortcoming because the structural topics are not available in the politician.json dataset

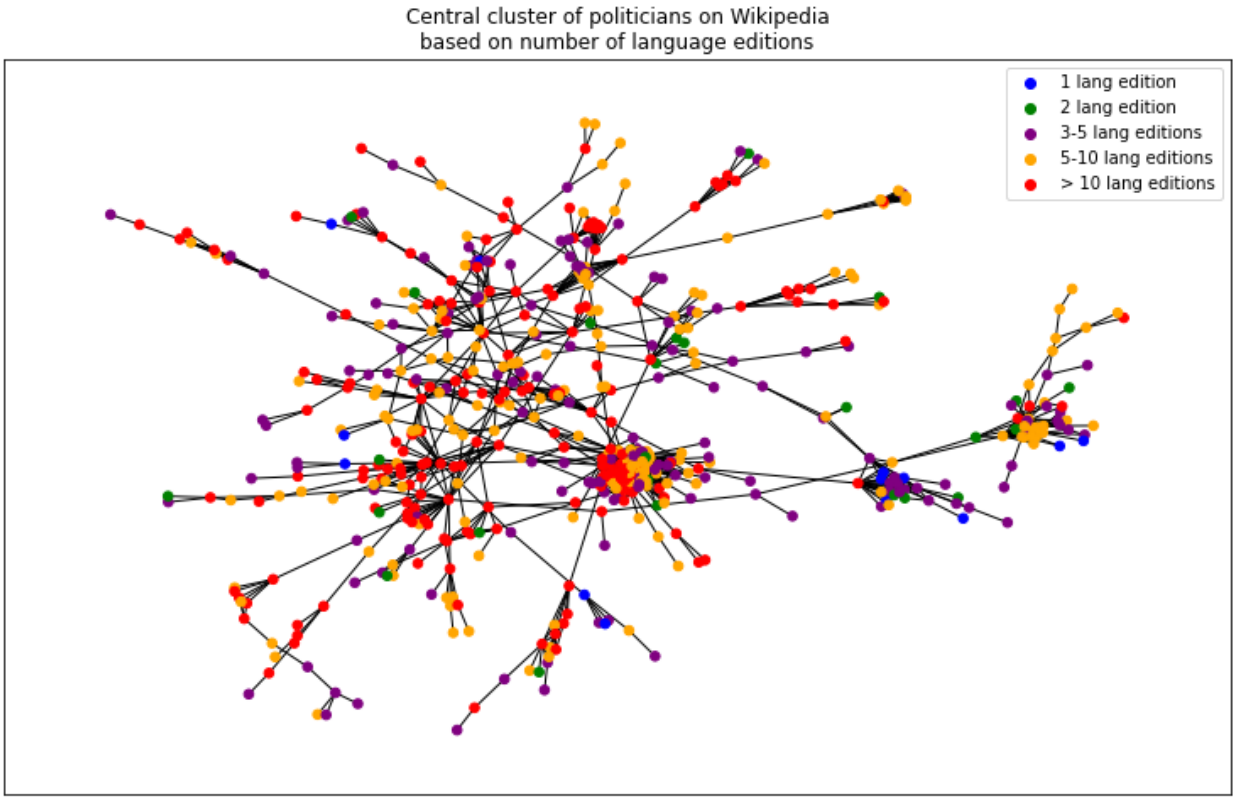
4. Visualize the network between the politicians based on gender and number of language editions. In this task, I plot the politicians network and identify the cluster. Then I proceed to plot the central cluster in the middle of the graph, which forms a global network of politicians.





We can see there is a clear central cluster of politicians, which form a global structure. Bigger images of the cluster after removing the peripheral politicians at the edges are shown below





Based on this visualization, it is sure that the closer a politician to the center of this cluster becomes, the more likely they are to be male and covered in a large number of language editions.

5. Regarding the central nodes, I use two different criteria for measuring the centrality: degree centrality based on the number of edges for each node, and closeness centrality, which measures how close a node is with regards to the rest nodes in the graph. After identifying the most central nodes in the network according to both criteria, these are some of my results:

- The central nodes are overwhelmingly male politicians, suggesting that the politic stage is dominated by men
- For the top central politicians, males usually have more language editions than females
- However, for the peripheral politicians, females are relatively more notable than their male counterparts and thus have more language editions

```
Number of male politicians out of top 1000 central ones by degree centrality: 873
Number of female politicians out of top 1000 central ones by degree centrality: 127
```

```
Average of language editions for top male central politicians by degree centrality: 10.348
Average of language editions for top female central politicians by degree centrality: 9.984

Average of language editions for bottom male central politicians by degree centrality: 6.106
Average of language editions for bottom female central politicians by degree centrality: 6.209
```

```
Number of male politicians out of top 1000 central ones by closeness centrality: 869
Number of female politicians out of top 1000 central ones by closeness centrality: 131
```

```
Average of language editions for top male central politicians by closeness centrality: 10.452
Average of language editions for top female central politicians by closeness centrality: 9.794

Average of language editions for bottom male central politicians by closeness centrality: 5.628
Average of language editions for bottom female central politicians by closeness centrality: 6.765
```

6. These are the observed differences between the politicians born before and after 1900

- There are much more documented male than female politicians pre-1900 compared to post-1900
- Politicians born before 1900 tends to have less cross references in the links, probably due to lack of information and unavailability of the internet, so relationships are harder to quantify
- However, politicians before 1900 are much more influential, so they have more language editions

```
Number of male politicians born before 1900: 967
Number of female politicians born before 1900: 27
Ratio of male/female politicians born before 1900: 35.81481481481482
Number of male politicians born after 1900: 1390
Number of female politicians born after 1900: 295
Ratio of male/female born after 1900: 4.711864406779661
Average number of language editions for politicians born before 1900: 5.914
Average number of language editions for politicians born after 1900: 3.985
Average number of links for politicians born before 1900: 2.055
Average number of links for politicians born after 1900: 3.227
Average number of politicians born before 1900 without death date: 0.048
Average number of politicians born after 1900 without death date: 0.806
```

7. The number of connected components with each distinct type of genders reveal the inherent bias in the editors of Wikipedia. We can see that in the original network, the number of only-male and only-female subgraphs are very high compared to the baseline random graph

```
Number of male only subgraphs in original graph: 174
Number of female only subgraphs in original graph: 5
Number of mixed gender subgraphs in original graph: 53
```

```
Number of male only subgraphs in baseline graph: 57
Number of female only subgraphs in baseline graph: 1
Number of mixed gender subgraphs in baseline graph: 9
```

These findings serve as evidence to support the claim in the paper where it states that female editors are much more interested in editing biographies about women politicians and vice versa for male editors. However, another limitation can be seen in the mixed gender subgraphs, as it is harder to prove whether this reduction is based on randomness of the baseline network or by the nature of the bias in editors.

8. I apply the same workflow from Task 2 and Task 3 for the artist.json and finns-1900-1940.json to see whether other nationalities and occupations have different biases than the current network

- Regarding linguistic bias, there is great emphasis on the personal life and relationship of female Finnish politicians compared to their male counterparts. On the other hand, career and professional language are much more dominant in articles about male Finnish politicians
- Finnish female politicians featured on Wikipedia, however, seem to enjoy the same notability as their male politicians.
- For artists, the biased language is less obvious, as the personal and professional term occurrence rates happen nearly equally for both male and female artists.
- However, there is a very large gap between the number of editions and the number of covering languages, where females are tremendously more notable than male artists (local heroes). This proves the claim in the paper that males are more self-absorbed than females and often initiate to write Wikipedia articles about themselves, especially in entertainment industries

```
Average occurence of personal terms for Finnish politicians: Male: 0.444, Female: 4.0
Average occurence of professional terms for Finnish politicians: Male: 1.519, Female: 0.143
```

```
Mean number of language editions for male Finnish politicians: 19.111111
Mean number of language editions for female Finnish politicians: 17.571429
Number of languages covering more Finnish female politicians than male politicians: 30
Number of languages covering more Finnish male politicians than female politicians: 33
```

```
Average occurence of personal terms for artists: Male: 0.818, Female: 1.069
Average occurence of professional terms for artists: Male: 0.214, Female: 0.19
```

```
Mean number of language editions for male artists: 8.399824
Mean number of language editions for female artists: 9.884726
Number of languages covering more female artists than male artists: 102
Number of languages covering more male artists than female artists: 46
```

3. Attached project code